早稲田大学博士論文(概要)

2011	学位記	文科省報告
	5970	甲 3616

SEMを利用した
新しい探索的データ解析法の開発

福中 公輔

目 次

第0章　本研究の位置づけ

本研究ではグラフィカル構造方程式モデリング（Graphical Structural Equation Modeling; 以下GSEMと略す）というデータ解析における新しい探索的なアプローチを提案する。このGSEMは，数学におけるグラフ理論を統計学のモデリングに応用した手法である。グラフ理論では統計学における変数を「頂点」，変数間の関係性「辺」として表現し，その集合を論じることになる。このようなグラフ理論を統計学へと応用する場合，変数として扱う対象（データ）の観点によって，質の異なる複数のモデリング・アプローチが存在する。ここでいう「観点」とは,「個人」なのか「組織・集団」なのか，あるいは「顕在変数」なのか「潜在変数」なのかのことである。グラフ理論の統計学への応用に関する発展の流れを図1に示す。

グラフ理論を利用したモデリングの基本は,「個人」を対象にした社会ネットワーク分析（social network analysis ; 以下SNAと略す）と考えられる。これは個人間の相互的関係性をグラフによって図示するための分析手法である。SNAを扱った例としては，学校教員がクラス内の子どもたちの友人関係をモデル化したものが有名であるが，個人相互の情報が必要となるため，データが得にくく，扱える範囲がかなり限定されている。

このSNAの発展系として，複雑ネットワーク（complex networks）がある。この分析が扱う範囲は，主に組織や集団であり，巨大な集団そのものをネットワークでモデル化し，その性質を論じるための分析手法である。使用する頂点数は数千から数万の規模のものが多い。近年の研究成果ではスモールワールド・ネットワークやスケールフリー・ネットワークが有名である。

SNAや複雑ネットワークは，個人間の相互関係性をモデル化できるという点で

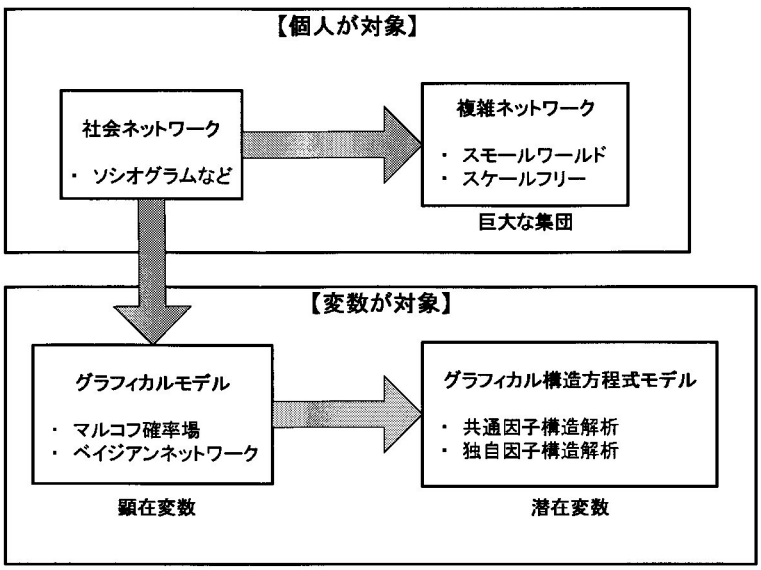

図 1: グラフ理論の統計的モデリングへの応用

有用ではあるが，個人がネットワークに所属する全員のデータを有していなければいけないという点で，データの収集がかなり困難であるという欠点を持つ。そこで個人の情報を変数としてまとめ，その変数間の情報をネットワークに要約すれば，データを手軽に収集でき，かつネットワークの規模自体が小さくなるため結果の解釈が容易になるというメリットがある。このような観点から見たモデリング技法が,「マルコフ確率場」と「ベイジアンネットワーク」である。マルコフ確率場は変数間の関係性を無向グラフで表したものであり，またベイジアンネットワークは変数間の関係性を有向グラフで表現したモデルである。これら 2 つの分析を合わせて，グラフィカルモデル（graphical model; 以下 GM と略す）と呼ぶ。

この GM は個人を縮約し，変数間の関係性をモデル化できるので，確かに有用ではあるが，人文科学系，特に心理学の扱う領域では使用されることが少なかった。なぜなら，心理学では実際には観察できない心理特性という対象を扱うので，顕在変数のみを扱った GM が活用できる状況が少なかったためである。したがって，この GM が潜在変数にまで使えるようになれば，さらに便利である。このような観点から開発した分析手法が，本研究で提案した GSEM である。GSEM は確

認的因子分析モデルにおける共通因子を対象にした共通因子構造解析と，独自因子を対象にした独自因子構造解析によって構成されている。

このように本研究は，グラフ理論の統計学への応用に関して，発展的系譜の一端を担うものであり，その研究意義は非常に大きいと考えられる。

第1章　SEMにおける探索的データ解析

探索的データ解析（Exploratory Data Analysis，以下EDAと略す）とはその名の通り，調査あるいは実験などで得られたデータの構造を探索的に探り出すことを目的とした統計的データ解析手法の総称である。EDAはデータの平均や相関など基礎統計量の考察から，多変量解析による分析や分析結果の視覚化（visualization）に至るまで，その扱う範囲は幅広い。また最近の動向の1つとして，Webシステムや POSシステムなどにより，自動的に集められた大量のデータから，意味のある情報の取得を目的としたデータマイニング（data mining）という解析手法が注目されているが，これもEDAの一種である。

本章ではこのEDAの特徴を振り返りながら，SEMでEDAを実践するための先行研究について考察した。まず始めにEDAの特徴として，渡部・鈴木・山田・大塚（1985）により重要と言われている「抵抗性」「残差の分析」「再表現」「図示」の4つの基本的な概念について解説した。次に探索的因子分析の詳細をまとめ，それをSEMに融合させた近年最も新しい分析手法の1つである探索的構造方程式モデリング（Exploratory SEM；以下ESEMと略す）と呼ばれる手法の考察を行った。その際，実データを用いてESEMに適用し，その有効性を検証した。

ESEMはSEMに探索的因子分析を融合することで，本来的には確認的分析であるはずのSEMをモデル探索にも使用できるという点で優れている。しかし，探索可能な空間は因子パタンに属するところのみであり，潜在変数間のパスの探索や誤差変数間の探索には使用できない。またESEMが前提としている状況は，因子数は既知であるが因子の意味は未知であるような場合であるが，このような条件

が成立するような状況はごくまれであろう。ESEM は飽くまで修正指標を利用する代わりのモデル修正のための方法として考案されたということである。したがって，SEM を EDA 的なアプローチとして使いたい場合，これらの点を解決し，さらには機械的な探索と同時に自身の研究仮説も反映できる探索法を考案する必要がある。

第2章　機械学習における探索的分析

統計学とは異なる分野で発展してきたデータ解析法の1つに機械学習（machine laerning）というものがある。計算機科学の分野を起源とする機械学習は，もともとは人間が有している学習能力と同等の機能を計算機に実装するための手段として，主に人工知能（artificial intelligence）の分野で研究がなされてきた。しかし研究が進むにつれて，人工知能だけではなく，大量のデータから計算機を用いて自動的に有効な知見を発見するためのデータ解析として用いられるようになった。近年では，これはデータマイニングとして知られている。

上記でも述べたが，データ解析における機械学習の大きな特徴は，与えられたデータから自動的に有効な知見を発見する目的で使用されることである。なお，ここで言う有効な知見とは「重要なデータの抽出」「異常データの検出」「データ間の関連性の探索」「モデルの発見」「規則の発見」などのことである。そして発見した知見を基にして，それらの結果を一般化し，新規データに対して予測及び制御を可能にすることが求められる。これら一連の流れが機械学習の扱う範囲である。

本章では，EDAにおいて使用される，代表的な機械学習の手法について解説を行う。具体的にはグラフィカルモデル（graphical model；以下GMと略す）を取り上げ，それらの手法の仕組みや意義を詳細に検討した。

GMは有向グラフのみで表現されたベイジアンネットワークと無向グラフのみで表現されたマルコフ確率場の2つに大きく分けることができる。このうちベイジアンネットワークは変数間の影響関係をモデル化できるので，そこから得られる知見はきわめて有効であるが，変数の増加に伴い，探索空間が爆発的に広がり，計算量が飛躍的に増大してしまう。共分散選択を利用することで計算量は減少す

るが，その場合，すべての変数間に順序関係が必要となってしまうため，実践場面で有効に活用するためには現実的でない。

一方，マルコフ確率場は変数間の空間的対称性を明らかにでき，変数の直接的な関係性を論じる上できわめて有効な知見を得ることができる。特に，人文社会科学系で取り扱うような，変数の数が少ないような場合には，共分散選択を利用することで簡便にモデル探索を行うことが可能である。

しかし，ベイジアンネットワークとマルコフ確率場の双方に共通して言えることは，どちらも観測変数のみに対してしか対応していないということである。心理学をはじめとした人文社会科学系の分野では，直接的には観測できないものを扱う必要があるという性質から，潜在変数を利用してモデルを構成することがよくある。そのため，GMが心理学系の研究分野で利用されることは少なかった。この点，すなわち潜在変数に対してモデル探索を行うことができないという点が，GMにおける最大の問題点である。

第3章　グラフィカルモデルのSEMによる表現

本章の目的は，

1. SEMにおける共分散構造の枠組みでGMの数理的アイデアが表現できることを理論的に示すこと
2. SEMによる分析結果の数値が，日本品質管理学会テクノメトリックス研究会（1999）の数値例に一致することを確認すること
3. 当該ソフトウェアのマニュアルには載っていない実行方法を解説し，実践の便宜に供すること

の3点である。

GMにおいて「モデル比較」というと，観測変数間にパスが引けるか引けないかという，共分散選択に基づくモデルの比較を意味している。しかし統計学には，因子分析モデルやパス解析モデルを始めとした，様々な統計モデルが存在するわけであり，これらの各種モデルとGMによるモデルとを比較することはきわめて重要であると考えられる。つまり，あるデータにGMを適用するのなら，そのモデルが他の統計モデルよりも最適であると判断できる根拠が必要であるということだ。

しかしこれまでの多くの研究では，これら各種モデルとGMとを客観的指標に基づいて，正しく比較されることはなかった。これは，そもそもGMという分析手法が多くの尤度モデルに基づく分析手法と根本的に異なるものであると考えられていたため，仕方のないことでもある。

しかし本章での研究により，GMもSEMの枠組みで解くことができ，構造方程式モデルの下位モデルに含めることが可能となった。これはすなわち，当該データに対してGMを適用するのが良いのか，あるいは因子分析モデルやパス解析モデルなどを適用するのが良いのかを，適合度という客観的基準を用いて，同じレベルで比較可能になったということである。

また，GMのもう一つの特徴として，これまでの研究ではほとんどの場合において観測変数のみに対してしか適用されてこなかったというものがある。しかし，本研究によってSEMと融合されたことにより，潜在変数に関してもより柔軟にGMを適用できる可能性が開けた。このように観測変数を主な対象としていたGMを，それと同レベルの水準で潜在変数への適用の可能性を広げたことは，応用可能性という面で非常に有用だと考えられる。

第4章　共通因子構造解析

心理学の研究には現在のところGMが頻繁に用いられているとは言いがたかった。その理由の１つとして，心理学の研究では，１つ１つの変数の信頼性が必ずしも高くない多数の観測変数を分析することが多いという事情が挙げられる。観測変数が幾つかの構成概念を測定するグループに分かれている場合には，GMは必ずしも有効に機能しないことが多く，このような場合には，もし因子分析によって抽出された因子に関して，GMのマルコフ確率場が適用できたら有用である (廣野・林，1994，小島，2003)。ただし，単純に因子間相関行列に直接GMを適用することは望ましくない。何故ならばGMは因子間相関を変化させるので，抽出された因子自体が変質するからである。この場合は飽くまでも因子の推定と因子間のGMを同時に推定し，データとの適合を吟味しなければならない。

これらのことを考慮し，本章では因子分析と因子間のGMを同時に行う共通因子構造解析と呼ぶ方法を提案する。適用例では，この方法を用いて因子間の相互関係を明らかにし，偏相関とクリークを基にして，因子間の対称的関連を考察する。

本研究における最初の２つの適用例では，因子分析研究において有名な相関行列を用いて，因子の抽出とそのGMを同時に行った。その結果，ここで利用した２つの先行研究が，人も時代も全く異なっているにもかかわらず，類似した構造を確認でき，知能全体に関する１つの仮説を獲得するに至った。一方３つ目の適用例に関しては，不適解が発生したために共分散選択が途中で停止し，完全な形で因子間の相互関係を明らかにすることができなかった。しかしこの結果は，今後の研究の土台として役に立つ，因子間関係の暫定的な構造になりうると考えられる。

一般的にGMのマルコフ確率場モデルにおける辺の有無は，通常のSEMにおけ

る双方向パスの有無には直接的には対応しない。しかし，本研究で開発した手法でGMを実施することにより，因子間の関連構造のより深い考察に役立つ情報を得ることができる。これが本研究においてマルコフ確率場モデルをSEMの枠組みに導入した最大の特徴である。

GMはデータから探索的にモデルを構築していく手段として非常に優れた手法である。これは研究仮説に基づくとは言え，ある程度恣意的にパスを引く必要があるSEMにはないモデル構成手段である。もちろん心理学の研究において研究仮説の重要性を否定するものではない。しかしその仮説の正しさをデータの側から検証できるGMは，モデル探索という観点からSEMにとって非常に有効な手段の1つになると考えられる。このような点から見てもGMをSEMの枠組みで表現したことの意義は大きい。

第5章　独自因子構造解析

因子分析においてモデルの改良を行う場合，因子数や因子パタンを変更することがよくある。しかしこれらのアプローチは，研究当初の構成概念の意味を変更，あるいは歪曲してしまうという問題点がある。またSEMの観点から独自因子に相関を仮定して対処することも可能であるが，どの因子間に相関を仮定すべきかの基準は曖昧である。そこで本研究では，共分散選択によるGMを利用して，研究当初に仮定した構成概念の意味を再解釈せず，かつデータからの明確な理由でモデルを改良するための独自因子構造解析という新しい探索的アプローチを提案する。

これまで，因子分析モデルの改良に関して，因子数を変更したり，因子から観測変数へのパスの数を変えたりする方法は多くの研究で散見されてきた。しかしこれらの方法を採用した場合は，仮定した構成概念の意味自体が変質してしまうため，抽出した因子の再解釈が必要となる。これにより，研究初期に仮定したモデルとは大きく異なったモデルが得られる場合もあるだろう。このため研究の方針や目的を変更せざるを得なくなる可能性がある。

ただし，モデルの試行錯誤を経て新たな問題に気づいたり，優れた着想を得ることもあるので，モデルを大きく変更することがすべての場合において誤りというわけではない。しかし研究している問題の種類によっては，当初のモデルを大きく変更することなく分析を続行したい，あるいはそうせざるを得ないこともある。このような場合に本研究で提案したアプローチは非常に有用である。

また，モデルの改良においては，SEMの観点から独自因子間に双方向のパスを引いて対処するアプローチが取られることもある。しかしこの場合，先行研究による知識も自身の研究仮説もなく，単に適合度を上昇させるという目的のためだ

けに恣意的になされることがほとんどである。これはSEMで確認的に分析する場合において，最もしてはならない対処法の1つである。しかし本研究で提案した方法はGMの共分散選択を利用してモデルの改良を行っているので，恣意的に行うよりも遥かに説得力がある。

一方，分析の結果に基づいてモデルの考察を行うとき，これまでは共通因子の解釈や因子パタンについては深く考察されてきたが，独自因子に対する考察はあまり行われてこなかった。これは独自因子が，「共通因子では説明できなかったその他すべての要因」という解釈しづらい性質を帯びているためだと考えられる。しかし本研究で提案した独自因子構造解析という手法は，独自因子間の直接的な関係を見いだすことが可能なため，共通因子では説明できない要因とその関連構造の発見のためにも，ひいては新しい因子を発見するという目的にも利用することが可能である。

このため本手法は，SEMにおける新しい探索的分析法となりうる。このように本研究で提案したアプローチは，モデルの改良や独自因子における考察，そして新しい知見の発見といった，データ解析における多くの状況で利用できる極めて有効な手法であるといえる。

第6章　総合考察

本論文では，一貫して機械学習におけるGMを統計学におけるSEMに融合するための方法論及び数理的枠組みの提案を行ってきた。これは確認的データ解析の用途として用いられることが多かったSEMに対して，探索的データ解析として使用するためであり，SEMにおける新しいアプローチであると言える。

本論文ではまず共分散選択を利用した古典的なGMをSEMの下位モデルとして表現した。これにより観測変数のみを使用したGMは，SEMのソフトウェアを用いて完全に分析可能であることが明らかとなった。次に，GMをSEMの枠組みで表現できたことを利用して，潜在変数間に対してもGMが適用可能であることを示した。ここでは確認的因子分析モデルにおける共通因子に対して共分散選択によるGMを実行し，共通因子間の構造を探索することを目的としている。その結果，研究仮説が少ないような場合でも，素早く共通因子間の関係性を明らかにすることができ，かつ優れた考察や気付きを得ることができた。ここで提案した方法を共通因子構造解析と呼ぶ。

最後に，潜在変数の中でも特に独自因子に対してGMを適用するための方法論を提案し，よりよいモデルを構築するための構造探索に利用する手段を開発した。この手法を独自因子構造解析と呼ぶ。独自因子構造解析では共通因子の構造を維持したままモデルを改良するための方法を提供することを目的としている。しかし，独自因子間に強い構造が見受けられるなど，初期モデルが著しく不適な場合などには，どのようにモデルを修正すべきかの見通しを与えてくれる。このように独自因子構造解析は，モデル修正のための副次的な効果も得られる極めて柔軟な探索的解析法である。

これら3つの分析法は，それぞれに独自の役割があり，分析のための目標を持っている。例えば，通常のGMは観測変数間のみの構造の探索を目的としており，潜在変数間の構造探索には活用できない。一方，共通因子構造解析は，潜在変数の中でも特に共通因子間の構造探索に使用でき，共通因子の相互の関係性から研究対象に対する深い考察と洞察を目標としている。また，独自因子構造解析は，潜在変数の中でも特に独自因子間の構造探索に活用でき，初期モデルを維持したままモデルの改良を行ったり，あるいは初期モデルの修正における方向性の把握を目標としている。

これらの手法はそれぞれが有用ではあるが，3つの手法を相互に組み合わせることでさらに有益な分析体系として確立することができる。この3つの手法を一連の流れとしてまとめた，SEMおよびGMを利用した探索的分析法の手続きがグラフィカル構造方程式モデリング（GSEM）である。GSEMを利用した探索的分析の一連のフローチャートを図6.1に示す。

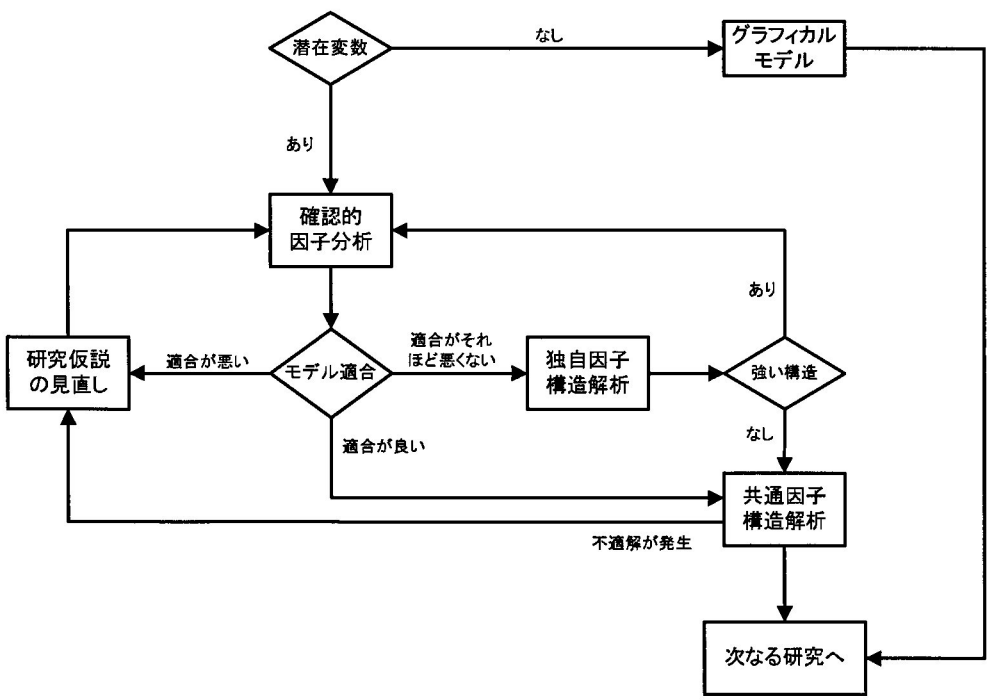

図 6.1: GSEMの手順

まず分析の出発点となるのは，モデルに潜在変数があるか否かである。もし潜在変数が存在しなければ古典的なグラフィカルモデルを用いて分析すれば良い。そして分析も結果を考察し，次なる研究へとつなげていくことになる。

しかし，自分が扱いたいモデルに潜在変数が存在した場合，古典的グラフィカルモデルは使用できない。その場合はまず確認的因子分析を行い，構成概念の意味を確定させることから分析を始める。確認的因子分析の結果，もしモデルの適合が悪ければ質問項目の見直し，場合によっては研究仮説そのもを見直す必要が出てくるだろう。また，適合がそれほど悪くなかったならば，独自因子構造解析を行い，独自因子間の構造を考察する。このとき独自因子間にクリークが構成され，強い構造が見られたならば，今度は因子数を変更するなどして確認的因子分析をやり直し，構成概念を再度検討することになる。いずれにせよ，ある程度適合が良い確認的因子分析モデルが確定するまでこの分析を繰り返す。その結果，構成概念が確定すれば，次のステップとして共通因子構造解析を行う。そして，その結果を考察しながら，次なる研究へとつなげ，最終的な構造方程式モデルを完成させていくことになる。これがGSEMというSEMの探索的なアプローチである。